Corrigé 9

Exercice 1. (a) Notons p le pourcentage recherché, et considérons $p \in (0,1)$. Si on choisit par hasard une personne parmi les étudiant-e-s de l'EPFL, celle-ci sera une femme avec la probabilité p et un homme avec la probabilité $1-p$. On peut définir une variable aléatoire

$$X = \begin{cases} 1 & \text{si la personne choisie est une femme,} \\ 0 & \text{si la personne choisie est un homme.} \end{cases}$$

La loi de cette variable est $\mathcal{B}(p)$.

(b) Le paramètre d'intérêt est p.

(c) Puisqu'il serait difficile d'observer toutes les personnes qui étudient à l'EPFL, on va observer un sous-ensemble. Ce sous-ensemble doit être représentatif, par exemple on peut observer un certain nombre d'étudiant-e-s qui mangent dans une grande cafétéria pendant la pause de midi.

(d) Un choix intuitif est le pourcentage de femmes dans le sous-ensemble observé.

(e) Même si on connaissait la valeur de p, on ne connaîtrait pas en avance la valeur de l'estimateur. Si l'on va dans la même cafétéria deux jours différents et l'on observe le même nombre d'étudiant-e-s, ce ne seront pas exactement les mêmes étudiant-e-s, donc on n'obtiendra pas le même résultat.

(f) On suppose que $p = 0.4$ et $n = 100$. D'après la partie (a), on peut supposer que les observations $x_1, \dots, x_{100}$ constituent une réalisation de $X_1, \dots, X_{100} \overset{iid}{\sim} \mathcal{B}(p)$. L'estimateur proposé dans la partie (d) s'écrit $\hat{p}_{100} = \bar{X}_{100} = (\sum_{i=1}^{100} X_i)/100$.

$$\begin{aligned} \mathbb{E}[\hat{p}_{100}] &= \mathbb{E}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) = \mathbb{E}[X_1] = p, \\ \mathrm{Var}[\hat{p}_{100}] &= \mathrm{Var}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) = \frac{1}{100}\mathrm{Var}[X_1] = \frac{p(1-p)}{100}, \\ \mathrm{b}(\hat{p}_{100}) &= \mathbb{E}[\hat{p}_{100}] - p = 0. \end{aligned}$$

L'estimateur $\hat{p}_n$ est non-biaisé. Si la taille de l'échantillon augmente, la variance diminue. Donc, avec un plus grand échantillon, on estime le pourcentage avec une plus grande précision (on s'attend à être plus proche de la vraie valeur).

(g) Remarquons tout d'abord que nous sommes ici dans la même situation que dans l'Exercice 1 de la Série 8. Les variables $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées, d'espérance $\mu = p$ et de variance $\sigma^2 = p(1-p)$. Nous pouvons donc utiliser le théorème central limite pour approximer la loi de

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} = \sqrt{n}\,\frac{\hat{p}_n - p}{\sqrt{p(1-p)}}.$$

Pour trouver n tel que $\mathbb{P}(\hat{p}_n < 0.5) \geq 0.95$ on calcule (avec $p = 0.4$)

$$\begin{aligned} & \mathbb{P}(\hat{p}_n < 0.5) \geq 0.95 \\ \Rightarrow\ & \mathbb{P}\left(\sqrt{n}\,\frac{\hat{p}_n - 0.4}{\sqrt{0.4 \times 0.6}} < \sqrt{n}\,\frac{0.5 - 0.4}{\sqrt{0.4 \times 0.6}}\right) \geq 0.95 \\ \Rightarrow\ & \Phi\left(0.204\,\sqrt{n}\right) \geq 0.95 \\ \Rightarrow\ & \sqrt{n} \geq \frac{\Phi^{-1}(0.95)}{0.204} \\ \Rightarrow\ & n \geq 65.42. \end{aligned}$$

Donc on a besoin d'observer au moins 66 personnes.

Exercice 2. (a) Les variables X_i sont discrètes, donc la fonction de vraisemblance est

$$L(p) = f_1(x_1; p) \times f_2(x_2; p) \times \ldots \times f_n(x_n; p),$$

où $f_i(x_i; p) = P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$ est la fonction de fréquences pour chaque X_i. On trouve

$$L(p) = p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\ldots p^{x_n}(1-p)^{1-x_n} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

(b) L'estimateur des moindres carrés est la valeur de p qui minimise

$$S(p) = \sum_{i=1}^{n}(x_i - p)^2.$$

Pour trouver une telle valeur on résout d'abord l'équation $S'(p) = 0$:

$$\begin{aligned} & S'(p) = 0 \\ \Leftrightarrow \quad & 2\sum_{i=1}^{n}(x_i - p) = 0 \\ \Leftrightarrow \quad & \sum_{i=1}^{n} x_i = n\,p \\ \Leftrightarrow \quad & p = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n. \end{aligned}$$

Il faut maintenant vérifier qu'il s'agit bien d'un minimum. On remarque que la fonction $S(p)$ est en fait un polynôme quadratique en p dont le coefficient de p^2 est strictement positif. Plus précisément,

$$S(p) = \sum_{i=1}^{n} x_i^2 - 2\,p\sum_{i=1}^{n} x_i + n\,p^2.$$

Donc la seule valeur p telle que $S'(p) = 0$ est le minimum global de la fonction. Par conséquent, $\bar{X}_n$ est bien l'estimateur des moindres carrés, $\hat{p}_{MC} = \bar{X}_n$.

(c) L'estimateur du maximum de vraisemblance est la valeur de p qui maximise $L(p)$, ou, de manière équivalente, la valeur qui maximise la fonction $\ell(p) = \log(L(p))$.
On a

$$\ell(p) = \sum_{i=1}^{n} x_i \log(p) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p).$$

Pour trouver le maximum on résout

$$\begin{aligned} & \ell'(p) = 0 \\ \Leftrightarrow \quad & \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0 \\ \Leftrightarrow \quad & (1-p)\sum_{i=1}^n x_i - p\left(n - \sum_{i=1}^n x_i\right) = 0 \\ \Leftrightarrow \quad & \sum_{i=1}^n x_i = pn \\ \Leftrightarrow \quad & p = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}_n. \end{aligned}$$

Il s'agit bien d'un maximum, étant donné que

$$\ell''(p) = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{n - \sum_{i=1}^n x_i}{(1-p)^2} < 0,$$

quel que soit $p \in (0,1)$. Donc la valeur $p = \bar{x}_n$ maximise la fonction $L(p)$ et $\bar{X}_n$ est l'estimateur du maximum de vraisemblance, $\hat{p}_{ML} = \bar{X}_n$.

(d) On a $\hat{p}_{MC} = \hat{p}_{ML} = \bar{X}_n$. Donc

$$\mathbb{E}[\hat{p}_{MC}] = \mathbb{E}[\hat{p}_{ML}] = \mathbb{E}[\bar{X}_n] = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = p,$$

parce que les variables X_i sont toutes $\mathcal{B}(p)$. Donc les estimateurs sont non-biaisés. Pour la variance on a

$$\begin{aligned} \mathrm{Var}[\hat{p}_{MC}] = \mathrm{Var}[\hat{p}_{ML}] = \mathrm{Var}[\bar{X}_n] = \\ = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{p(1-p)}{n}, \end{aligned}$$

parce que les variables X_i sont indépendantes et toutes $\mathcal{B}(p)$.

Exercice 3. (a) On sait que $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. Donc

$$1 = \int_0^1 c\,x^{\theta-1}\,\mathrm{d}x = c\left[\frac{x^\theta}{\theta}\right]_0^1 = \frac{c}{\theta},$$

et on voit bien que $c = \theta$. On a donc la densité

$$f(x) = \begin{cases} \theta\,x^{\theta-1} & \text{si } x \in (0,1) \\ 0 & \text{sinon.} \end{cases}$$

(b) On a

$$\mathbb{E}[X_1] = \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x = \int_0^1 x\,\theta\,x^{\theta-1}\,\mathrm{d}x = \theta\int_0^1 x^\theta\,\mathrm{d}x = \theta\left[\frac{x^{\theta+1}}{\theta+1}\right]_0^1 = \frac{\theta}{\theta+1}.$$

(c) Les variables X_i sont continues et on note $x_1, \ldots, x_n$ leurs réalisations. Ainsi, la fonction de vraisemblance est

$$L(\theta) = f_1(x_1;\theta) \times f_2(x_2;\theta) \times \ldots \times f_n(x_n;\theta),$$

où $f_i(x_i;\theta) = f_i(x_i)$ est la densité pour chaque X_i. On trouve

$$L(\theta) = \theta\, x_1^{\theta-1} \theta\, x_2^{\theta-1} \ldots \theta\, x_n^{\theta-1} = \theta^n \left(\prod_{i=1}^n x_i\right)^{\theta-1}.$$

Donc

$$\ell(\theta) = \log(L(\theta)) = n\log(\theta) + (\theta-1)\sum_{i=1}^n \log(x_i).$$

Pour trouver la valeur de θ qui maximise $\ell(\theta)$ on résout

$$\begin{aligned} & \ell'(\theta) = 0 \\ \Leftrightarrow \quad & \frac{n}{\theta} + \sum_{i=1}^n \log(x_i) = 0 \\ \Leftrightarrow \quad & \frac{1}{\theta} = -\frac{1}{n}\sum_{i=1}^n \log(x_i) \\ \Leftrightarrow \quad & \theta = -\frac{n}{\sum_{i=1}^n \log(x_i)}. \end{aligned}$$

Il s'agit bien d'un maximum puisque

$$\ell''(\theta) = -\frac{n}{\theta^2} < 0,$$

pour tout $\theta > 0$. Donc la valeur $\theta = -\frac{n}{\sum_{i=1}^n \log(x_i)}$ maximise la fonction $L(\theta)$ et $-\frac{n}{\sum_{i=1}^n \log(X_i)}$ est l'estimateur du maximum de vraisemblance, $\hat{\theta}_{ML} = -\frac{n}{\sum_{i=1}^n \log(X_i)}$. Remarquons que puisque $x_i \in (0,1)$, on a $\log(x_i) < 0$ et par conséquent $-\frac{n}{\sum_{i=1}^n \log(x_i)} > 0$.

Exercice 4. On sait que le numéro le plus élevé dans le canton va être au moins aussi grand que le plus grand numéro observé dans le stationnement. On va donc estimer le numéro le plus élevé dans le canton par un numéro $m \geq 298158$. Si on prend $m = 298158$, on "sait" intuitivement qu'on va sous-estimer. On voudrait prendre $m > 298158$, mais si on va "trop loin" de 298158, on va sur-estimer. Dans l'exercice suivant on va voir comment on peut choisir l'estimateur pour que, en moyenne, on ne sous-estime ou ne sur-estime pas (cela veut dire que si on répète la même expérience beaucoup de fois, l'espérance de l'estimateur va être le numéro cherché).

Exercice 5. (a) Les variables X_i sont continues, donc la fonction de vraisemblance est

$$L(\theta) = f_1(x_1;\theta) \times f_2(x_2;\theta) \times \ldots \times f_n(x_n;\theta),$$

où $f_i(x_i;\theta) = f_i(x_i)$ est la densité pour chaque X_i. Pour la loi uniforme $\mathrm{U}[0,\theta]$ on a la densité

$$f(x) = \begin{cases} 1/\theta & \text{si } x \in [0,\theta] \\ 0 & \text{sinon.} \end{cases}$$

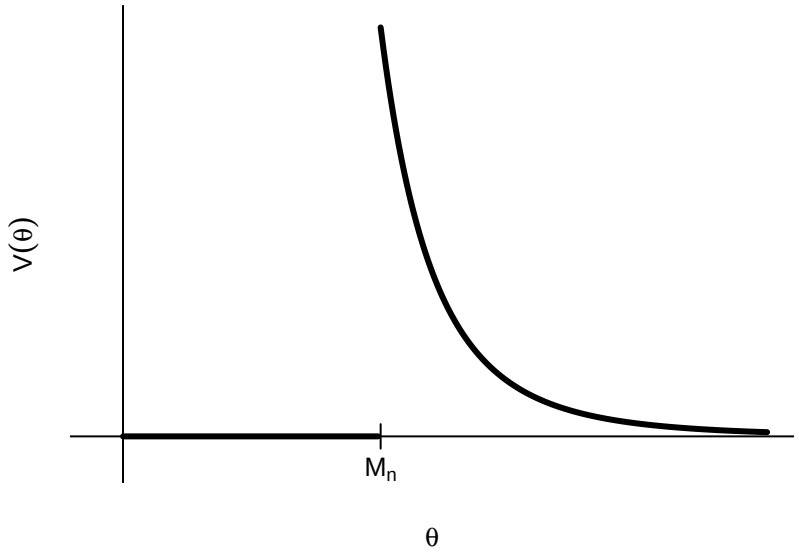
La fonction de vraisemblance est donc

$$L(\theta) = \begin{cases} 1/\theta^n & \text{si } x_i \in [0,\theta] \text{ pour } i \in \{1,\dots,n\} \\ 0 & \text{sinon.} \end{cases}$$

Autrement dit,

$$L(\theta) = \begin{cases} 1/\theta^n & \text{si } \max_{i\in\{1,\dots,n\}} x_i \leq \theta \\ 0 & \text{sinon.} \end{cases}$$

(b) Avec $M_n = \max(X_1,\dots,X_n)$, le graphe de $L(\theta)$ est



(c) On voit sur le dessin que cette fonction est maximale pour $\theta = M_n$. Notons que $L(\theta)$ n'est pas dérivable, donc le maximum ne peut pas être trouvé en utilisant $\ell'(\theta)$ comme dans les exercices précédents.

(d) Pour trouver le biais de M_n il faut calculer $\mathbb{E}[M_n]$. M_n est une variable aléatoire continue, donc $\mathbb{E}[M_n] = \int_{-\infty}^{\infty} x f_{M_n}(x)\,\mathrm{d}x$, où $f_{M_n}(x)$ est la densité de M_n.
Pour trouver cette densité, calculons d'abord la fonction de répartition de M_n. On a, pour tout $x \in [0,\theta]$,

$$\begin{aligned} F_{M_n}(x) &= P(M_n \leq x) = P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) \\ &\overset{(1)}{=} P(X_1 \leq x)P(X_2 \leq x)\cdots P(X_n \leq x) \overset{(2)}{=} P(X_1 \leq x)^n \\ &= \left(\frac{x}{\theta}\right)^n, \end{aligned}$$

où on utilise l'indépendance des variables pour (1) et le fait qu'elles sont identiquement distribuées pour (2). D'autre part $F_{M_n}(x) = 0$ pour $x < 0$ et $F_{M_n}(x) = 1$ pour $x > \theta$. En dérivant $F_{M_n}(x)$ on trouve la fonction de densité voulue :

$$f_{M_n}(x) = \begin{cases} n\,\frac{x^{n-1}}{\theta^n} & \text{si } x \in [0,\theta] \\ 0 & \text{sinon.} \end{cases}$$

Maintenant on peut calculer

$$\mathbb{E}[M_n] = \int_{-\infty}^{\infty} x f_{M_n}(x)\,\mathrm{d}x = \int_0^\theta n\,\frac{x^n}{\theta^n}\,\mathrm{d}x = \left[\frac{n}{n+1}\,\frac{x^{n+1}}{\theta^n}\right]_0^\theta = \frac{n}{n+1}\theta.$$

L'estimateur M_n est donc biaisé : son biais est

$$\mathrm{b}(M_n) = \mathbb{E}[M_n] - \theta = -\theta/(n+1),$$

donc M_n sous-estime θ (ce qui est en accord avec notre intuition).
Pour obtenir un estimateur non-biaisé on pose $\hat{\theta}_{NB} = \frac{n+1}{n} M_n$. Sa variance est $\mathrm{Var}(\hat{\theta}_{NB}) = (\frac{n+1}{n})^2 \mathrm{Var}(M_n)$.
Pour trouver $\mathrm{Var}[M_n]$ on calcule d'abord

$$\mathbb{E}[M_n^2] = \int_{-\infty}^{\infty} x^2 f_{M_n}(x)\,\mathrm{d}x = \int_0^\theta n\,\frac{x^{n+1}}{\theta^n}\,\mathrm{d}x = \left[\frac{n}{n+2}\,\frac{x^{n+2}}{\theta^n}\right]_0^\theta = \frac{n}{n+2}\theta^2.$$

Donc

$$\begin{aligned}
\mathrm{Var}(M_n) &= \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \frac{n}{(n+1)^2(n+2)}\theta^2, \\
\mathrm{Var}(\hat{\theta}_{NB}) &= \frac{(n+1)^2}{n^2} \times \frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{1}{n(n+2)}\theta^2.
\end{aligned}$$

(e) On peut demander que l'estimateur soit non-biaisé. Dans ce cas, on choisira $\hat{\theta}_{NB}$ puisque $\hat{\theta}_{ML}$ est biaisé.
On peut demander que l'erreur quadratique moyenne soit la plus petite possible. On a

$$\begin{aligned}
\mathrm{EQM}(\hat{\theta}_{ML}) &= \mathrm{Var}(\hat{\theta}_{ML}) + \mathrm{b}(\hat{\theta}_{ML})^2 = \frac{n\theta^2}{(n+1)^2(n+2)} + \frac{\theta^2}{(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)}, \\
\mathrm{EQM}(\hat{\theta}_{NB}) &= \mathrm{Var}(\hat{\theta}_{NB}) + \mathrm{b}(\hat{\theta}_{NB})^2 = \frac{1}{n(n+2)}\theta^2.
\end{aligned}$$

On obtient donc facilement que

$$\mathrm{EQM}(\hat{\theta}_{ML}) - \mathrm{EQM}(\hat{\theta}_{NB}) = \frac{\theta^2(n-1)}{n(n+1)(n+2)}.$$

Ainsi, pour tout $\theta > 0$ et $n \geq 2$, l'EQM de $\hat{\theta}_{NB}$ est strictement inférieure à celle de $\mathrm{EQM}(\hat{\theta}_{ML})$. C'est donc de nouveau l'estimateur $\hat{\theta}_{NB}$ qui est préférable.

(f) On trouve $M_n = 298158$ et $\hat{\theta}_{NB} = 313065.9$.